Evaluating Multi-modal Input Modes in a Wizard-of-Oz Study for the Domain of Web Search

Alexandra Klein, Ingrid Schwank, Michel Généreux & Harald Trost

Austrian Research Institute for Artificial Intelligence,
Schottengasse 3, A-1010 Vienna, Austria
Tel: *+43 1 5324621 1*
Fax: *+43 1 5336112 77*
Email: *{alexandra,michel,harald}@ai.univie.ac.at*
URL: *http://ai.univie.ac.at*

Browsing documents in the World Wide Web usually offers little opportunity for real interaction, as user input is often limited to mouse clicks and to filling out forms. By adding language technology, the gap between navigation and interaction can be bridged, which should lead to a more symmetric communicative setting. It means that free typed or spoken utterances can be used as means of access to documents on the Web, in addition to mouse clicks and typed strings as they are used to fill out forms, e.g. in search engines.

A crucial factor in system design is a sound pre-design study, which investigates how users will use typed and spoken queries to complement existing functionality. In this paper, we are going to describe a Wizard-of-Oz (WoZ) study which has shown that users tend to prefer multi-modal input over traditional input modes in search tasks concerning German newspaper texts, and that they were generally faster in completing the tasks whenever they were allowed to use free typed and spoken input. Additionally, users were grouped according to previous experience with search engines and the Internet. It can be shown that non-expert users expressed a stronger preference for multi-modal interaction than expert users and that, with multi-modal interaction, task completion times were reduced more significantly among non-experts than among experts users.

Generally, we interpret these effects as an indication that multi-modal access to the Web may result in higher usability, especially for non-expert users.

Keywords: natural language processing, multi-modality, usability, Wizard-of-Oz experiments.

1 Introduction

While developers of Web pages may use any combination of text, audio, pictures and video in their presentations in order to address the user effectively — thus fully exploiting the multimedia possibilities of the Web — the users' reactions are much more limited in being restricted mainly to point-and-click operations. It is to be expected that users can overcome this communicative asymmetry better if they are offered a greater variety of input modes, namely multi-modal interaction including free typed and spoken utterances. In the project which is described in this paper, we concentrate on the role of multi-modal queries posed to a search engine from the point of view of usability and its implications for the system architecture and the interface.

Recently, first steps have been taken by augmenting the Web by search engines which accept words or even phrases entered by the user and which are then matched with the stored information. We argue that these input options are still far too formal and too different from natural conversations to achieve high user acceptance once more and more interaction is carried out on the Web.

There are also structural arguments for using free spoken or written language. Complex types of action cannot be handled by mouse clicking and typing simple phrases, as these two types of input usually only refer to visible entities. Therefore, it is not possible to express e.g. indirect relations occurring throughout natural speech. furthermore, simple reference modes like single-word search or point-and-click are often particularly inadequate as a lot of information on the Web is textual which suggests also textual access. By using the latter, the gap between navigation and interaction in a communicative setting can be bridged. Textual queries also provide the advantage of reaching through the hypertext structure directly to the required (textual) information. This frees the user from a dependence on the document structure offered by the content provider, which is advantageous because often, users' and content providers' intentions may differ. At the same time, the user is not restricted to any predefined wording of the query.

2 Multi-modal access to texts on the Web

In order for the user action to be interpreted correctly, the constraints, which are imposed by the communicative setting, need to be recognised and represented. This suggests an action-centred model which treats user action, be it by spoken or written input or by point-and-click operations, as instances of information requests in the specific communicative context. This way, the expressive power of user requests is acknowledged by not restricting the user like in command-and-control systems.

At the same time, regarding user action in its situational contexts helps reduce the ambiguities and errors which are introduced by the results of speech recognition (Oviatt, 2000b).

Research and development in speech recognition has recognised the advantages and challenges of multi-modal interaction. System design has mostly focused on integrating various input modes, paying special attention to speech as the most relevant one. In these systems, browsing functionality can be accessed by appropriate speech commands. This approach is sensible, e.g. for offering additional browsing functionality over the telephone and for providing a restricted speech recognition vocabulary, resulting in better system performance. Yet these systems are still far from a real integration of reference to form, content and structure, as it is natural for any interaction in the context of hypertext documents. Therefore, it is the aim of the described project to point out new ways of integrating speech and language with classical access methods, and to investigate the respective shortcomings and advantages of different combinations of input modes. In contrast to commercial systems, not a broad solution will be reached. Rather, a situation with complex interaction is created and analysed in terms of usability, while the domain is limited in this research scenario.

In order to gain insight into the possibilities of speech access to the Web, from a practical point of view, the project is concerned with building the prototype of an interactive system for Web browsing and search, which integrates the processing of spoken and typed queries with the usual point-and-click commands of a GUI.

While the functionality of the GUI part will be standard, natural language queries can concern any combination of the following:

- what some particular piece of information refers to (i.e. content);
- how this information is presented (i.e. structure); and
- how this information is connected (i.e. document hierarchy).

Accordingly, we expect the system to handle (the German equivalents of) spoken browser commands, e.g. *back*, as well as content queries, e.g. *more about soccer*, and combinations of the two, e.g. *back to the page about the soccer game*. Natural language queries are to be interpreted in the appropriate contexts as they can be derived from the current communicative setting. The language processing module distinguishes queries referring to content, structure or document hierarchy and determines the new appropriate system states, e.g. by displaying text segments.

In the system architecture, on the other hand, the functionality of the GUI is enhanced and complemented by the possibility to use language commands. It seems plausible to stipulate that adding spoken and typed natural utterances to the input media of a Web browser will generally lead to a higher usability. This hypothesis was supported by the Wizard-of-Oz experiments which were carried out in the pre-design study. Therefore, it is very important to integrate the three input modes speech, text and point-and-click operations in a way that is perceived as natural by the user, in order to achieve high usability. Thus, the contrastive functionality (Oviatt & Olsen, 1994) of the input modes concerning interaction with the system is considered.

3 The Wizard-of-Oz Study

3.1 Aims

In the context of the project, the pre-design study aims at three main goals:

- research on the role of speech input in a multi-modal system;
- the adaptation to user needs; and
- the creation of a corpus.

A prominent aim in usability testing is the investigation of mode preference and coordination. Generally, researchers agree that a contrastive functionality of several input modes including speech guarantees highest performance as well as user acceptance (Oviatt & Olsen, 1994). The experiments are designed to investigate how system functionality is addressed by the input mode and which function is fulfilled by speech with respect to the other variants.

The investigation parallels research on speech input as opposed to other input modes in the domain of map tasks (Cohen et al., 1994), dialogue agents (Litman, 1998) and multi-modal access to timetable information (Quarfordt, 1998).

Experience with multi-modal interfaces integrating speech has generally shown that only a cyclic development process results in a usable system, cf. (Mariani, 1997). At several stages of the development process, user responses have to be considered and evaluated.

Furthermore, it is another important aim of usability testing in the proposed project to collect a corpus for user interaction in the specific domain of German search requests for news texts on the Web.

It was decided to use Wizard-of-Oz experiments (Fraser & Gilbert, 1991) in the pre-design study. Wizard-of-Oz experiments imply that the test persons are given tasks for interaction with a system, but they are not aware of the fact that the system is at least partly simulated by human(s), the *wizard(s)*. Wizard-of-Oz experiments are often used in the development of Natural Language Processing systems because with them, it is possible to test functionality, user behaviour and interfaces for systems which have not yet been fully implemented. The results of the Wizard-of-Oz tests can then be incorporated into the design of the actual systems.

We are well aware of the problems associated with WoZ experiments in a speech environment, particularly the simulation of speech recognition errors (Aust et al., 1994; Oerder & Aust, 1994) and possibly the lack of relation to real-world tasks. Yet, testing the usability of the interface and obtaining a corpus is an important impulse for the development of the system. For the experiments, we decided to simulate almost perfect speech understanding. Naturally, errors in speech understanding and recognition are an important aspect in user interaction with the system (and in usability). Yet, it is very difficult to adequately simulate speech understanding errors, particularly speech recognition errors, and consequently, we decided not to model this part of system behaviour.

3.2 The Setup of the experiments

The four performance dimensions in the evaluation are task completion (determined in retrospect), task completion speed (which will be measured), complexity of interaction (derived from logging the interaction) and the user response to the system (evaluated by means of questionnaire-based satisfaction ratings (Litman, 1998)).

The subject's utterances were recorded, and typed input and mouse clicks were logged in order to assess task complexity. These sources can be synchronised via time stamps. Variance among the performance dimensions is evaluated according to standard methods of empirical evaluation (Preece et al., 1994).

The interface was built using Netscape Navigator 4.73 as a browser. The interface consists of two frames. One frame was simply used for displaying the resulting pages as provided by the more important control frame. To avoid the activation of links in the display frame by the user, the control frame had to intercept all sorts of unwanted events that could accidentally be triggered by the user. This was achieved by the use of an *applet*, a Java program associated with the page. Because the applet had to cope with all sorts of events which are usually handled naturally by the browser, and also because we wanted to limit user interference, the editor was made very simple. The applet was also responsible for triggering a commercial speech synthesis tool. It uses the Java Speech Application Programmer Interface, which allows for smooth interaction with commercial speech products. The applet was also responsible for producing a log file containing the user's and wizards' keyboard actions (the wizard's actions were limited to displaying Web pages and error messages, both triggered by key combinations).

The Wizard-of-Oz effect was produced by another computer connected to the same network as the computer used by the test person. A freely available program called PC Remote allowed the wizard to view and control the user's interface via a set of keys. The test persons' spoken utterances were recorded using a free software called STX, from the Austrian Academy of Sciences.

3.3 Test persons

43 test persons participated in the experiments. Before the start of the tests, they filled out a questionnaire asking for information on previous experience with search engines and the Web in general. According to the degree of familiarity with search engines and the Web, users were grouped as *experts* and *non-experts*. Familiarity was measured with the help of questions concerning expertise with traditional search engines, mainly concerning knowledge and use of logical operators. As is has been postulated that especially non-expert users benefit from multi-modal interaction (Oviatt, 2000a), for a Wizard-of-Oz study in our context, these two users groups were evaluated comparatively in addition to the comprehensive evaluation of all users. Overall, experts as well as non-experts had various professional backgrounds and belonged to all age groups between 18 and 62.

3.4 Tasks

Each test person was assigned 12 tasks. In each task, one newspaper text had to be searched in a large databases of German newspaper texts. The documents looked like texts on the Web but were stored locally on our server and where not changed

	general reaction	dialogue	interface	overall
non-experts	1.6	1.8	1.9	1.8
experts	2.1	2.3	2.4	2.3
average	1.9	2.1	2.1	2.0

Table 1: User ratings concerning general topics.

	System 1 spoken + mouse	System 2 spoken + written + mouse	System 3 written + mouse
non-experts	1.77	1.90	1.93
experts	2.00	2.13	2.34
average	1.88	2.01	2.10

Table 2: User ratings concerning systems.

during the course of the experiments. In the specification of the tasks, each text was summarised in two sentences, and the test persons had to formulate adequate queries based on these summarisations.

The twelve tasks were grouped according to combinations of input modes, during the course of the experiments, all tasks appeared with all input modes. We called the three different combinations 'systems'. One system accepted only spoken input and mouse clicks, one system accepted only typed input and mouse clicks, and one system accepted spoken and typed input as well as mouse clicks.

Each test person was given an instruction sheet, but was advised to use free spoken and typed input. The interface was explained, but no demonstrations were given. Assistance was available on demand, but only help concerning the interface or formal issues was given. Error messages were displayed whenever test persons mixed up the systems or formulated requests completely irrelevant to the tasks. This way, test persons were discouraged to explore the limits of the 'system' as this would have led to problems for the evaluation.

After completion of all tasks, the test persons filled out another questionnaire. This questionnaire focused on usability of the system as a whole as well as the different combinations of input modalities. This way, it is possible to compare users' preferences with task completion times under different constraints for the input modalities.

3.5 Results

Overall, the test persons stated that they consider the system usable (cf. Table 1). On a scale between 1 (highest grade) and 5 (lowest grade), the mean grade is 2.049. Generally, non-experts liked the system more than experts. This supports the hypothesis that the combination of traditional input modes with free typed and spoken input is appealing particularly to non-expert users, as it is a more natural way of interaction (cf. Table 2).

	System 1 spoken + mouse	System 2 spoken + written + mouse	System 3 written + mouse
non-experts			
average number of interactions	24	22	22
average time in seconds	553	514	1000
experts			
average number of interactions	26	25	26
average time in seconds	593	427	758

Table 3: Number of interactions and time used.

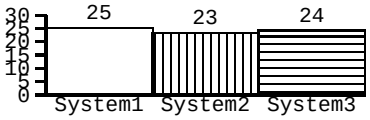

Figure 1: Number of interactions (mean).

Regarding specific aspects of usability of the system, the results can be compared to the overall grades given to the system. Efficiency received the grade 2.05 in overall average.

Furthermore, it can be said that on average, test persons needed less time for the completion of the task if they were allowed to use spoken or spoken and written language. Generally, task completion times were longer if the test persons only used written input (cf. Table 3). Expert users improved less by the use of spoken language than non-experts users did. For System 1 (spoken input and mouse clicks), there is no significant difference between non-expert and expert users. For System 2 (spoken and written input as well as mouse clicks), there is a significant difference. The same can be said for System 3. Generally, expert and non-expert users were faster using all three input modes (cf. Table 1 and Table 2).

So far, results indicate that multi-modal interaction can provide a usable and efficient access to documents on the Web, especially for non-experts. Further analysis of the experiments will focus more on the role of free spoken and written input. It has to be determined which functions these input modes tend to fulfill: whether they are fall-back modes, a general alternative or means of getting a shortcut to the requested information. In our tests, only a fraction of users replaced mouse clicks with spoken commands. Some extremely unexperienced users, however, relied solely on spoken commands and chose not to use mouse clicks. Search commands were usually spoken rather than typed. There was a significant growth in competence for the users during the tests. Again, these results show that using multi-modal interaction is an efficient means of accessing the Web, particularly for non-expert users.

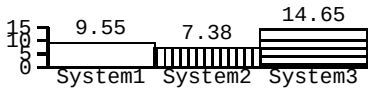

Figure 2: Number of minutes (median).

4 Conclusion

The Wizard-of-Oz study has shown that users, particularly non-expert users, prefer multi-modal interaction over traditional written input modes and point-and-click operations as they are used in Graphical User Interfaces, at least in the domain of searching newspaper texts and under idealised speech recognition conditions. Furthermore, it can be shown that users tend to perform faster if they are able to use spoken-language queries, and also then, non-expert users had a greater gain in task completion times. This is crucial as documents in the Internet have become an important source of information in both professional and private contexts, and access should be as efficient, open and natural as possible, not only from a technical but also from a communicative point of view.

The results from the Wizard-of-Oz experiments are considered in the design of the system, and users' comments will influence a new version of the interface. While there are results of Wizard-of-Oz experiments e.g. for multi-modal access to timetables (Quarfordt, 1998), there is no sound empirical foundation for usable interaction schemes accessing pages in the Web. Empirical evidence from this domain will contribute to research on multi-modal interfaces and speech understanding for different applications. From the point of view of natural language processing, the obtained empirical data has two interesting dimensions. On the one hand, the corpus will provide important insights into the structure of communication, in this case the interaction with the Web as a multi-modal environment, and how it is mirrored in user action (speech, typing, point and click). On the other hand, from a more application-oriented point of view, empirical results can provide a valuable link between natural language understanding components and objectives of software usability.

Acknowledgements

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture; the project is supported by the Austrian Science Fund (P13704-INF). The authors would like to thank the users who participated in the study as well as the anonymous reviewers for their comments.

References

Aust, H., Oerder, M., Seide, F. & Steinbiss, V. (1994), Experience with the Philips Automatic Train Timetable Information System, *in* ***EDITOR*** (ed.), *Proceedings IVTTA'94 the 2nd IEEE Workshop on Interactive Voice Technology For Telecommunications Applications*, IEEECSP, pp.67–72.

Cohen, P., Johnston, M., McGee, D., Oviatt, S., Clow, J. & Smith, I. (1994), The Efficiency of Multimodal Interaction: A Case Study, *in* ***EDITOR*** (ed.), *Proceedings the International Conference on Spoken Language Processing (ICSLP'94)*, Acoustics Society of Japan, pp.249–52.

Fraser, N. M. & Gilbert, G. N. (1991), "Simulating Speech Systems", *Computer Speech and Language* **5**(1), 81–99.

Litman, D. L. (1998), Evaluating Response Strategies in a Web-based Spoken Dialogue Agent, *in* ***EDITOR*** (ed.), *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol. II, Université de Montreal, Morgan-Kaufmann, Montreal, Quebec, Canada, pp.780–6.

Mariani, J. J. (1997), "Spoken Language Processing and Multimodal Communication: A View from Europe", Plenary Talk, NSF Workshop on Human-centered Systems: Information, Interactivity, and Intelligence (HCS), Arlington, VA, USA.

Oerder, M. & Aust, H. (1994), A Real-time Prototype of an Automatic Inquiry System, *in* ***EDITOR*** (ed.), *Proceedings the International Conference on Spoken Language Processing (ICSLP'94)*, Vol. 2, Acoustics Society of Japan, pp.703–6.

Oviatt, S. L. (2000a), Multimodal Interface Research: A Science without Borders, *in* B. Z. Yuan, T. Y. Huang & X. F. Tang (eds.), *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP2000/Interspeech 2000)*, China Military Friendship Publishers, p.***PAGES***.

Oviatt, S. L. (2000b), "Taming Recognition Errors with a Multimodal Interface", *Communications of the ACM* **43**(9). Special Issue on Spoken Language Interfaces.

Oviatt, S. L. & Olsen, E. (1994), Integration Themes in Multimodal Human-Computer Interaction, *in* ***EDITOR*** (ed.), *Proceedings the International Conference on Spoken Language Processing (ICSLP'94)*, Vol. 2, Acoustics Society of Japan, pp.551–4.

Preece, J., Rogers, Y., Sharpe, H., Benyon, D., Holland, S. & Carey, T. (1994), *Human–Computer Interaction*, Addison–Wesley.

Quarfordt, P. (1998), Usability of Multimodal Timetables: Effect of Different Levels of Domain Knowledge on Usability, Master's thesis, Linköpings Universitet, Sweden.

Author Index

Keyword Index